

SENTIMENT ANALYSIS OF 10-K REPORTS: TO WHAT EXTENT DO WE NEED SYNTACTIC INFORMATION?

BRAM GUBBELS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

u547289

COMMITTEE

Raquel G. Alhama
prof. dr. Zamberlan

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2022

ACKNOWLEDGMENTS

I would like to thank my supervisor assistant professor Alhama for her valuable support and guidance during the writing of procces of my thesis.

SENTIMENT ANALYSIS OF 10-K REPORTS: TO WHAT EXTENT DO WE NEED SYNTACTIC INFORMATION?

BRAM GUBBELS

CONTENTS

1 Introduction 4
2 Literature Review 6
2.1 Dictionaries 7
2.2 Advanced NLP Techniques in Finance 8
2.2.1 BERT 9
2.2.2 FinBERT 9
2.2.3 Recursive Neural Tensor Network 10
3 Method 10
3.1 Textual Data 10
3.2 Financial Data 11
3.3 Pre-processing 12
3.4 Model Comparison 12
3.5 Event Study: Stock Forecasting 13
3.6 Descriptive Statistics 15
4 Results 17
4.1 Financial Phrasesbank and FIQA 17
4.2 Event Study: Management Discussion and Analysis 23
5 Discussion 24
5.1 Contributions, limitations and suggestions for future research 26
6 Conclusion 27
7 Appendix A 31

LIST OF FIGURES

Figure 1 Visualization of the Methodology 15
Figure 2 Sentiment Loughran McDonald period 1999-2021 17
Figure 3 Sentiment FinBERT period 1999-2021 17
Figure 4 Accuracy of the four models on the datasets 19
Figure 5 F1-score of the four models on the datasets 19
Figure 6 Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from Financial Phrasebank 50% agreement. 20
Figure 7 Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the FIQA headline dataset. 21
Figure 8 Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the FIQA microblogs dataset. 31
Figure 9 Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the Financial Phrasebank 100% agreement. 32
Figure 10 Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the Financial Phrasebank 75% agreement. 33
Figure 11 Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the Financial Phrasebank 66% agreement. 34
Figure 12 Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the MDA section dataset 35
Figure 13 Sentiment RNTN period 1999-2021 35

LIST OF TABLES

Table 1 Summary Statistics of the variables: Sentiment LM, Sentiment FinBERT,Sentiment RNTN, CAR[-1,1], VIX, Length, ROE, Book-to-Market 16
Table 2 Correlation Matrix 17
Table 3 Model performance on Financial Phrase Bank, FIQA and MDA dataset. 18
Table 4 Estimation of the regression equation for the sentiments; depedent variable CAR[-1,1] 23

Abstract

The use of NLP techniques has gained wide recognition in the fields of finance and accounting during the last decade. However, despite the widespread adoption researchers rarely adopt the most advanced NLP techniques. This can lead to misclassification because the traditional approaches fail to incorporate deeper semantic information. In this paper I aim to compare the methods traditionally used such as Harvard IV-4 and Loughran and McDonald dictionary with more advanced techniques such as FinBERT and a Recursive Neural Tensor Network (RNTN) to see if incorporating syntactic information leads to better results. After which I will use the best performing models to see if you can use sentiment analysis to predict stock market returns from the Management Discussion and Analysis (MDA) section of annual reports (10-ks). I find that on the labeled datasets the FinBERT model outperforms all other methods. The superior performance is attributable to using syntactic information in combination with that the model is specially designed for financial language. The Harvard IV-4 and RNTN are both not built for financial language and perform the worst on the labeled datasets. Secondly I apply the Loughran and McDonald dictionary, FinBERT and RNTN sentiment classifier on the MDA section from 10-ks. According to the efficient market hypothesis when information is released the stock price will immediately incorporate the information, therefore the sentiment from the MDA section can help us forecast stock prices. I do an OLS regression on the cumulative abnormal returns (CARs) of the release of the 10-k and the sentiment from the MDA section of S&P 500 companies for the period of 1999-2021. The Loughran and McDonald sentiment is not significant, the sentiment from the Loughran and McDonald sentiment can therefore not be used to forecast stock prices. Both the FinBERT classifier and the RNTN classifier have a significant relationship with the stock returns and could therefore be used to forecast stock prices. However, the relationship is so small that the trading based on either of these models will not result in positive returns due constraints such as transaction fees.

Keywords *: Sentiment Analyses, Syntax, Financial Natural Language Processing, Stock Market*

number of words *8719*

1 INTRODUCTION

In this project, we aim to extract the sentiment of the management from annual reports (10-ks) because the management has an information advantage compared to outside investors and forecast stock prices based on the sentiments using lexicon-based sentiment classifiers and more advanced classifiers which incorporate syntactic information. First, a bit of context is given. Then, the motivation behind the need to forecast stock prices using sentiments and why we need to compare different models. Finally, the research questions are stated.

All listed companies in the United States are obligated to publish information about the company multiple times per year. The annual report or 10-k is one of these obligated publications. A 10-k is a comprehensive summary from the performance of the company in the previous year and expectations for the future. The Security and Exchange Commission determines the rules for the 10-ks and is responsible for the oversight of the listed companies. When companies release a 10-k they release information to the public. According to the efficient market hypothesis from Fama (1970) when information is released the stock price will immediately incorporate this information so when positive information is released for example a profit windfall this will have a positive effect on the stock price. However, this only works when the market did not already anticipate the news. When this is the case, you can use the difference between the stock price before the release of information and after the release of information to approximate the value of the information. In this paper I will use this mechanism to measure if a positive sentiment from the management in the annual report leads to an increase in the stock price and vice versa (Fama, 1970).

The use of textual analysis has become an important part of the toolbox in finance and accounting research in the last decade. Textual analysis is for example used for social media analysis (Das & Chen, 2007) , fraud detection (Hoberg & Lewis, 2017; Purda & Skillicorn, 2015) and sentiment classification (Kearney & Liu, 2014). Despite the increasing use of text analytics in finance researchers rarely adopt advanced NLP techniques. This might be explained by the complicatedness of the newest NLP techniques. One of the major differences in this study compared to previous research is that in addition to the use of dictionary-based models I will also use more advanced techniques such as the BERT model or a recursive neural tensor network (RNTN). I will use these models to see if using more advanced models which incorporate syntactic information improves the ability to capture the correct sentiment.

Loughran and McDonald (2016) created a survey of all available research regarding textual analysis in Finance and Accounting and concluded that the current available research lacks a combination of textual analytics and an economic hypothesis. In my research I will combine textual analytics and an economic hypothesis by researching if sentiment analysis can be used to forecast stock prices.

In this paper I evaluate the performance of different NLP models on the classification of sentiments for financial and accounting literature. For the first task I will test the accuracy and F1-score of four models: Harvard Iv-4 dictionary, Loughran and McDonald dictionary, FinBERT and a RNTN from the Stanford CoreNLP packages. The models are tested on three different labeled datasets, the Financial Phrase bank, Financial Opinion Mining and Question Answering (FIQA) dataset and 600 randomly selected sentences from the MDA section of 10-ks. The Financial Phrase bank consists of headlines about companies from the Finish OMX exchange. The FIQA database consists of microblogs posts from social media such as twitter and sentences from financial news articles.

RQ1 *To what extent can the level of syntactic information the models are capable of representing improve the accuracy of the sentiment classification of labeled financial sentences?*

For the second part of the paper, I will extract the sentiment from 10-k reports of the companies listed in the S&P 500 with the three best performing models from the previous research question. Then I will use an OLS regression with the cumulative abnormal returns (CARs) as a dependent variable and the sentiment as an independent variable. I will use this to see if there is a significant relationship between the CAR and sentiment and to see if there is a predictive power to the use of sentiment from the MDA section of an annual report. The regression consists of annual reports from 1999 to 2021 and includes more than 4000 observations.

RQ2 *: Does the sentiment extracted from the MDA of 10-k reports have a significant positive relationship with the abnormal returns surrounding the release of a 10-k?*

In this study I find that the FinBERT model outperforms the other three models; Harvard IV-4, Loughran & McDonald and the RNTN on the labeled datasets. Models who use a level of syntactic information and are designed to take into account the specific language in corporate filings outperform on the sentiment classification tasks. I therefore conclude that the use of advanced NLP techniques can improve the performance.

Furthermore, I find that the FinBERT classifier and the RNTN classifier have a significant relationship with the CARs of the S&P 500 companies for the period of 1999 and 2021. This would indicate that potential profitable trading strategies could be developed using these models, however when transaction costs are included the relationship is too small to lead to positive returns.

2 LITERATURE REVIEW

The use of textual analysis in finance and accounting has seen a large increase, this is likely because of two factors. First there has been an exponential increase in the computer power over the last 50 years and secondly the large increase in the availability of information. According to Li (2008) the introduction of the Electronic Data Gathering, Analysis and Retrieval system (EDGAR) – EDGAR is an online database with the required filings from all public listed companies - by the Securities and Exchange Commission (SEC) increased the accessibility and efficiency of corporate filings. The main research areas are targeted phrases, sentiment analysis and readability.

Most of the literature in Finance and Accounting looks to quantitative factors to explain stock prices. While there is no doubt that quantitative factors have helped to gain understanding of the reasons behind stock prices there is an increasing realization that quantitative factors alone might not be adequate to fully explain stock prices. Which is indeed supported by earlier work from Shiller (1980) and ROLL (1988) who showed that stock prices can not be explained alone by quantitative factors. It is however not surprising why the majority of the research focuses on quantitative data; it is easily accessible, more objective and easier to measure. The use of non-quantitative data is complicated by the lack of objective measurement techniques. However, the recent development in other research areas such as linguistics, Natural Language Processing and deep learning have made it easier to use non-quantitative data to explain stock market data. One of the first studies of the sentiment was from Feldman, Govindaraj, Livnat, and Segal (2010) who showed that non-financial information such as the sentiment extracted from the MDA of a 10-ks are significantly correlated with returns in a short-time window. Indicating that non-financial information contains relevant information for stock predictions. They used a method in which they compared the frequency of positive words and negative words using a dictionary specially designed for finance and accounting literature. The dictionary approach has been the standard approach in finance and accounting when classifying sentiment.

2.1 *Dictionaries*

The majority of the research was done by using one of these four dictionaries: Henry (2008) , Harvard IV-4, Diction and Loughran and McDonald (2011) . The Harvard IV-4 and the Diction dictionary were created not taking account the specific language used in finance and accounting literature. Li (2008) and Loughran and McDonald (2011) criticize the use of non-financial dictionaries to capture sentiment from corporate disclosures. According to Loughran and McDonald (2011) almost 75 percent of the negative words in the Harvard IV-4 do not have a negative meaning when used in a financial or accounting context. Words such as tax, cost, capital, board, liability and depreciation which occur often in corporate filings such as 10-ks and 10-Qs do not necessarily have a negative meaning. Also, industry specific words are likely misclassified for example in the pharmaceutical industry the use of diseases such as diabetes, alzheimer or cancer are considered negative however in the context of the pharmaceutical industry these are linked to cures in development and already existing cures which of course are not necessarily negative.

For the Diction dictionary Loughran and McDonald (2015) find that the vast majority of Diction words are misclassified. Frequently occurring Diction positive words such as respect, power, necessary and trust will not typically have a positive meaning when used in corporate disclosures and negative words like, no, not without, pain and gross tend not to mean something negative in corporate disclosures leading to the wrong classification.

The Henry (2008) dictionary is the first dictionary specially designed for corporate disclosures. The dictionary was created by examining the earnings press releases of the computer service industry and telecommunication companies. The dictionary consists of 105 positive words and 85 negative words. This is also the main caveat of the Henry (2008) dictionary which is the limited number of words. Where the Harvard IV-4 dictionary has a total number of 4100 negative words the Henry (2008) dictionary only has 85 negative words. However, Price, Doran, Peterson, and Bliss (2012a) find that at measuring the tone of earnings conference calls the Henry (2008) dictionary outperforms the Harvard IV-4 dictionary.

The Loughran and McDonald (2011) dictionary differs from the previous three because it has 6 different word categories: negative, positive, uncertainty, litigious, strong modal and weak model. The model was created by examining a sample of 10-k reports during the period 1994-2008. The dictionary consists of 353 positive and 2339 negative words. Compared to the other three dictionaries it has two main advantages: no commonly occurring negative and positive words appear to be missing and it was

designed specifically by examining 10-k reports. The Loughran and McDonald (2011) have become the most dominant of the dictionary, therefore I will use this dictionary for the dictionary-based technique however I will also compare the performance of the Loughran and McDonald (2011) dictionary with a non-financially designed dictionary the Harvard IV-4 to assess of the Loughran and McDonald (2011) also performs better when classifying multiple different sources of corporate disclosures.

The use of dictionaries to determine the sentiment of corporate filings also has caveats, according to Loughran and McDonald (2020) managers have the tendency to add a lot of positive words to negative statements. This to mask, hide or spin negative news. Kothari, Shu, and Wysocki (2009) find evidence for the idea that managers try to change how negative news is presented to the public; they find that managers withhold bad news from investors. Furthermore Cao, Jiang, Yang, and Zhang (2020) find that in the age of AI and machine learning where an increasing number of investment decisions are made by algorithms companies are preparing their corporate filings considering which words are considered positive and negative according to the dictionaries and adapt their filings. The manipulation and collision of the corporate filings can lead to misclassification. Although there is clear evidence that management adapts their filings based on the content of the dictionaries there might be a limitation on how extreme the management will change their filings due to potential liability by misinforming their stakeholders. According to Field, Lowry, and Shu (2005) the disclosure of bad news deters future litigations for the company and the management. The use of advanced NLP techniques may be able to overcome the disadvantages of the dictionary approach and improve the ability to correctly classify the sentiment.

2.2 *Advanced NLP Techniques in Finance*

Although the number of papers in finance and accounting research applying advanced NLP techniques is low there are papers who applied advanced techniques. One of the first papers which used deep learning methods for textual analysis in finance was from Kraus and Feuerriegel (2017).They used an LSTM neural network and found that it outperformed traditional methods. The LSTM network yielded considerably more information regarding future stock prices. There have been several others who also applied advanced techniques for sentiment classification to financial texts (Lutz, Pröllochs, & Neumann, 2019; Maia, Freitas, & Handschuh, 2018; Sohangir, Wang, Pomeranets, & Khoshgoftaar, 2017). Sohangir et al. (2017) applies several neural network models LSTM, word2vec and convolutional neural network. The models are applied on a stocktwits dataset which

is a forum where individual stocks are discussed. They find that deep learning can be used for financial analysis and that the convolution neural network is the best model to predict sentiment of stocktwits tweets. Maia et al. (2018) applies an LSTM network to classify sentences in the Financial PhraseBank. The initial works on sentiment classification in finance using advanced technique show that these models can be an improvement to the traditional methods. However, the progress of computational linguistics and NLP gives us the opportunity to use even more state-of-the-art models.

2.2.1 *BERT*

A large step forward in computational linguistics and NLP was the development of the BERT model by (Devlin, Chang, Lee, & Toutanova, 2018). BERT stands for Bidirectional Encoder Representations from Transformers. BERT is a multilayer bidirectional transformer encoder based on the transformer architecture Vaswani et al. (2017). The transformer model was first introduced by Vaswani et al. (2017) and is entirely based on attention replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self attention. The transformer model has two main advantages compared to RNN models. First, transformers have no locality bias because the input is processed all at once where RNN models lose information the longer a sentence is. Secondly, the transformer architecture has shorter training time due to the lower computation complexity and larger parallelization.

The model is pre-trained on BookCorpus (800 million) words and English Wikipedia (2500 million words). The BERT model has two variants BERT-base and BERT-large. BERT-base is modelled to the OPENAI GPT (Radford, Narasimhan, Salimans, & Sutskever, 2018) for comparison and has 12 layers and 110 million parameters. The BERT-large model has 24 layers and a total of 340 million parameters (Devlin et al., 2018). The fine-tuning of BERT is straightforward since the self-attention mechanism in the transform allows BERT to model many downstream tasks. For each task you plug in task specific input and outputs into BERT and fine-tune the parameters compared to pre-training fine-tuning is inexpensive. All of the results can be replicated in at most 1 hour on a single cloud TPU, or a few hours on a GPU, starting from the extracted pre-trained model (Devlin et al., 2018).

2.2.2 *FinBERT*

The BERT model can be easily fine tuned on a downstream task; however, it has been shown that pre-training large-scale domain corpora can improve the performance on prediction tasks (Beltagy, Lo, & Cohan, 2019; K. Huang,

Altosaar, & Ranganath, 2019; Lee et al., 2020). A. Huang, Wang, and Yang (2020) were the first to develop a BERT model pre-trained on specific financial communication. They compiled a large financial Corpora of 4.9 billion tokens consisting of 2.5 billion tokens from corporate reports 10-K and 10-Q, 1.3 billion tokens from earnings Call Transcripts and 1.1 billion from analyst reports. The FinBERT is then compared to the performance of the BERT model on sentiment classification tasks. The FinBERT model outperforms the BERT model at every task based on accuracy as performance measure. In financial markets where a small difference can already have a large impact, the usefulness of FinBERT to the BERT model seems evident. Araci (2019) compared the FinBERT model to LSTM, FinSSLX and a ULMFit model on the Financial PhraseBank dataset. The FinBERT model outperformed all other models on the accuracy and F1-score. However, the FinBERT has not yet been compared with the traditional used methods such as dictionaries and only to a small selection of advanced techniques.

2.2.3 *Recursive Neural Tensor Network*

Another advanced technique which as far as I am aware has not been applied to financial text classification is the Recursive Neural Tensor Network (RNTN) from Socher et al. (2013). In 2013 Socher et al. (2013) proposed a recursive neural tensor network in addition with a Stanford Sentiment Treebank. The treebank included 215,154 sentiment labels in the parse trees of 11,855 sentences and allows us to train and evaluate compositional models. The corpus of the Treebank consists of movie reviews. This is of course a disadvantage because movie reviews and finance and accounting language differ significantly from each other. It would be interesting to see if the ability to incorporate syntactic information can overcome the lack of financial base language and outperform the dictionaries designed for financial language. From a linguistic standpoint you would expect that including word order for semantic tasks will improve the classification because it is able to correctly classify sentences with negation.

3 METHOD

3.1 *Textual Data*

To answer the first research question the performance of the four models are examined on three datasets. The Financial Phrases databank (Malo, Sinha, Takala, Korhonen, & Wallenius, 2013), the Financial Opinion Mining and Question Answering dataset (FIQA) and MDA section dataset. The Financial Phrases databank consists of 4840 sentences from news articles

on companies listed on the Finish stock Exchange OMX Helsinki. The sentences were annotated by 16 experts with background knowledge of the financial markets. The dataset is divided into 4 subcategories, depending on the agreement of the experts, namely 50%, 66%, 75% and 100%. The second labeled dataset is the FIQA which originates from an open challenge in 2018 from the Web conference. The dataset consists of 675 microblog posts from reddit, twitter and wikinews and 438 financial headlines. The data is humanly-annotated and given a sentiment score between -1 and 1. I will transform the sentiment score to the labels positive, negative and neutral. A sentiment score below -0.5 is classified as negative where a sentiment score above 0.5 is considered positive and between -0.5 and 0.5 the sentiment score is classified neutral. The third dataset consists of 600 sentences randomly selected from the MDA section of S&P 500 companies during the period of 1999 to 2021. I have annotated the sentences myself with the labels: positive, neutral and negative. When annotating sentences from corporate filings knowledge from finance and accounting is required due to my previous education a master of Finance I am able to classify the sentences myself.

To answer the second research question, I obtained the annual report of the S&P 500 companies for the time period of 1999 until 2021. The data is gathered from the EDGAR filing system from the SEC which is a large dataset containing corporate filings of all US listed companies. The 10-k reports are gathered through a data scraper from Loukas, Fergadiotis, Androutsopoulos, and Malakasiotis (2021). The data scraper is used because the database from the SEC lacks an easily used interface to download large quantities of corporate filings. Furthermore, the data scraper removes tables and returns a cleaned text file. The S&P 500 consists of around 500 companies which represent the largest US listed companies. Due to bankruptcies, mergers, acquisitions, removal and adding of companies to the index the total number is not always equal to 500 and can deviate. Also, due to formatting issues and missing data the number of 10-ks analyzed is lower than the total number of companies in the S&P 500.

3.2 *Financial Data*

For the second research question to link the sentiment classification to the CARs stock market data is required. From the companies whose sentiment is classified the stock returns are gathered from a period of 1999 until 2021. The stock returns are gathered through the Wharton Research Data Services (WRDS) which is a data platform from the Wharton Business University for financial research. The data is gathered for a period of 1999 and 2021 and includes the ticker, closing price, shares outstanding,

the market return and the return of the individual companies. For the market return the S&P 500 index is used as a proxy. This is an average result of all the companies within the S&P 500 where the companies with a larger market capitalization and free float have a larger weight than smaller companies. Furthermore, the control variables for the OLS regression are book-to-market, return on equity (ROE), length of text and volatility are gathered through WRDS.

3.3 *Pre-processing*

First, the tables and HTML tags are removed so that we obtain a plain text file. Secondly, I remove all non-alpha characters and lower case the remaining text. Thirdly, sentence with less than 5 words such as ['see sources below'] and more then 64 words in it are removed because they are relative uninforming for the sentiment of the company and/or management. Finally double spaces are removed. After the pre-processing the text files are ready for processing using the sentiment algorithms.

3.4 *Model Comparison*

The labeled dataset described above will be used to answer the first research question, the pre-processed sentences will get a sentiment classification from four sentiment classification algorithms: Harvard IV-4, Loughran and McDonald, FinBERT and RNTN. The labeled data is used to first measure and compare the performance of the difference model before using the best performing models on the unlabeled data. The models are evaluated based on the accuracy and F1-score for each dataset separately. For the two dictionaries a sentence is considered positive when the number of positive words outweighs the number of negative words, a sentence is considered negative when the number of negative words outweigh the number of positive words, and a sentence is considered neutral when the number of positive words are equal to the number of negative words. The FinBERT model classifies a sentence positive-negative-neutral based on a SoftMax function. Lastly for the RNTN the sentences are classified very positive-positive-somewhat positive-neutral-somewhat negative-negative-very negative where I will transform this back to be in accordance with the other models to positive-neutral-negative where very negative and somewhat negative are transformed to negative and somewhat positive and very positive are transformed to positive. Secondly, I will examine the mistakes made by the different models and examine if syntactic information improves the performance and is able to solve problems such as negation. The first research question is used as a set up for the second in which I will

use the best performing models to see if I can use sentiment as a predictor for stock prices.

3.5 *Event Study: Stock Forecasting*

For the second research question I aim to measure the sentiment of the MDA section in a 10-k report and analyse if there is a relationship with stock returns. I will use the three best performing models on the labeled dataset. The MDA section of a 10-k report is divided into sentences and every sentence is classified as positive, negative or neutral similar to the previous section. To determine the sentiment for the entire MDA section I will follow the method developed by Henry and Leone (2016) and Price, Doran, Peterson, and Bliss (2012b) where the sentiment is defined as follows:

$$Sentiment = (Positive - Negative)/(Positive + Negative) \quad (1)$$

Where positive is the number of positive sentences per report and negative is the number of negative sentences per report. This will lead to a sentiment score between -1 and 1.

To determine the relationship between the stock returns and the sentiment extracted from the MDA section of the 10-k report I will do an Event Study. Event studies are a methodology used in the Finance and Accounting research to value the release of information (MacKinlay, 1997). Given rationality in the marketplace, the effect of the release of information in this case the sentiment from the management will be reflected immediately in the stock prices. Using a short window of time surrounding the release of information the economic effect of the release of information can be measured. This gives us the opportunity to understand if the market considers this information valuable and if it might be used for forecasting stock prices. To determine if the release of information leads to stock returns you can not simply take the stock returns on the day of the release of information but need to adjust potential other factors who could influence the stock return on this day. I will therefore use the method described by MacKinlay (1997) because according to a study from Cable and Holland (1999) who compared multiple techniques, they showed a strong preference for regression-based models where the market model based on the CAPM was preferred. I will explain this model briefly below, for more information you can read the article of MacKinlay (1997). Below you find the abnormal returns this is the stock return of a firm surrounding the release of information but adjusted for other factors that could have influenced the stock price.

$$AR_i t = R_i t - \beta_i * R_m t \tag{2}$$

Where AR is the abnormal return at time t of firm i, Beta i is the measure of systemic risk of a firm compared to the market portfolio and $R_m t$ is the return of the market portfolio. Below you find the Cumulative Abnormal returns which aggregate abnormal returns during the entire event window.

$$CAR_i(t1, t2) = \sum_{t=1}^{t2} AR_{it} \tag{3}$$

Where CAR is the cumulative abnormal return from t1 the start of the event window and t2 the end of the event window. The CAR is calculated for the event window (-1,1). To test if the CAR are significantly different from zero I will use an t-test, to see if the sentiment of the MDA section of a 10-k report has an effect on the stock return.

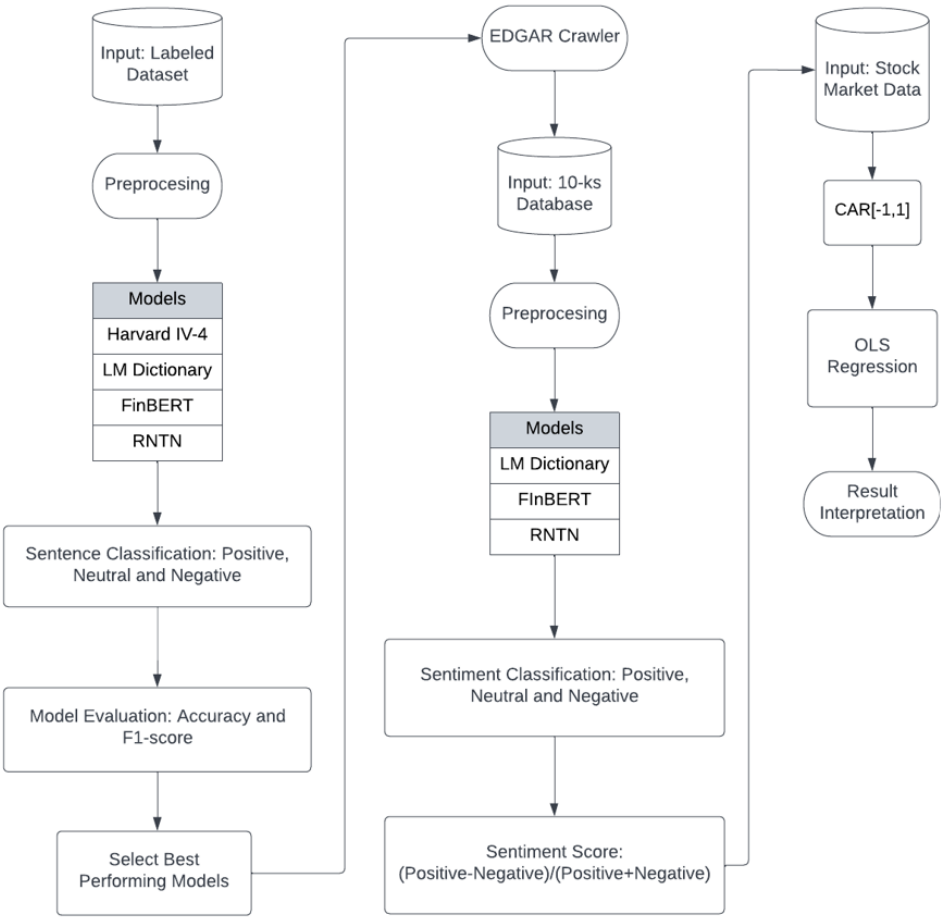
To determine the relationship between the dependent variable CAR[-1,1] and the independent variable sentiment I will do an OLS regression to determine the effect of sentiment on the stock returns surrounding the release of information. The short-window of [-1,1] is used because the smaller the window, the more accurate the returns can be measured. The event window starts a day before the release of the information due to potential foreknowledge. I include in addition to the sentiment a number of firm specific control variables to improve the robustness of the results. I will include as control variables the Return on Equity (ROE) of a firm in the period prior to the release of the 10-k report. The return on equity is a ratio of the net income divided by the equity of a company. The ratio is a measure of the profitability of a firm in relation to the outstanding equity. In addition to the ROE I will also include the volatility index of the S&P 500. THe volatility index is a measure of the expectations of the volatility from the Chicago Board Options Exchange which is determined based on the option market. The reason for including the VIX is that during periods of larger volatility due to uncertainty for example the response to releases of information can be larger, by including the VIX I will aim to control for this effect. Thirdly I control for the length of the text, the length variable counts the total number of sentences of a MDA section. Lastly I will include the book-to-market ratio, which is the book value divided by the market capitalization is a crude measure for if a company is expensive relative to the value on its books, where a high book to market indicates that the company is inexpensive and vice versa. The control variables are obtained through the compustat database on WRDS exempt for the VIX which is downloaded through the CBOE index database. From

the dependent and independent variable described below the following regression equation follows.

$$CAR[-1,1] = \beta_0 + \beta_1 * SentimentScore_t, i + \beta_2 * ROE_t, i + \beta_3 * Vix_t, i + \beta_4 textlenght_t, i + \beta_4 * book - to - market_t, i + errorterm \quad (4)$$

In figure 1 below you find a visualization of the methodology used in this paper.

Figure 1: Visualization of the Methodology



3.6 *Descriptive Statistics*

In table 1 the summary statistics of the variables are presented. The sentiment of the LM dictionary is overall negative with -0.2, where the sentiment from the FinBERT model is positive with 0.1053 and the sentiment from

the RNTN is negative with -0.209. The standard deviation of the LM dictionary is 0.22, for the FinBERT model 0.293 and for the RNTN 0.153. The dispersion of the sentiment is fairly similar for all three models. The FinBERT clearly finds a positive sentiment compared to the LM dictionary and RNTN. The CAR [-1,1] is 0.0004 indicating that there is a positive response surrounding the release of the 10-k, however it is only slightly above zero. The length of the MDA section is on average 427 sentences but can differ significantly with a maximum of 3592 and a minimum of only 5. Companies have a choice on how detailed and long the MDA section is. Furthermore , the average ROE is 18.5% with a standard deviation of 0.359. The Vix has a mean of 18.089 and a standard deviation of 5.98. The book-to-market has a mean of 4.8098 and a standard deviation of 21.539.

Table 1: Summary Statistics of the variables: Sentiment LM, Sentiment FinBERT,Sentiment RNTN, CAR[-1,1], VIX, Length, ROE, Book-to-Market

Variables	Mean	Std. Dev.	Min	Max	Observations
LM	-0.200	0.221	-0.833	0.727	4988
FinBERT	0.1053	0.293	-0.8570	0.95	4988
RNTN	-0.209	0.153	-0.8181	0.6129	4988
CAR [-1,1]	0.0004	0.0242	-0.282	0.212	4988
VIX	18.0894	5.979	9.14	76.45	4988
Length	472.707	378.259	5.0	3592.0	4988
ROE	0.185	0.359	-1.652	12.889	4416
Book-to-Market	4.8098	21.539	-376.94	990.205	4416

In table 2 the correlation between the multiple variables is presented. The FinBERT sentiment and the LM sentiment have a correlation of 0.507. The RNTN has a lower correlation to both the Loughran and McDonald sentiment and the FinBERT sentiment. All three sentiments are positively correlated with the ROE and the CAR[-1,1]. The number of sentences is positively correlated with the FinBERT and RNTN sentiment but negatively with the LM sentiment. The VIX negatively correlated with all three sentiments. The book-to-market ratio is negatively correlated to the LM and FinBERT, but positively to the RNTN. The level of correlation between the individual variables is low and does not indicate any problems with multicollinearity.

In figure 2, 3 and 13 the mean sentiment for each year of the included companies are plotted to see the overall sentiment change during the entire period of the dataset. The sentiment is normalized to get a mean of zero and a standard deviation of 1, this is done to make a better comparison between the two different sentiment classifiers. When plotting the sentiment you would expect that during economic downturns or stock market crashes the sentiment is lower and that during economic booms the sentiment is

Table 2: Correlation Matrix

	LM	FinBERT	RNTN	CAR	VIX	Lenght	ROE
LM	X	X	X	X	X	X	X
FinBERT	0.507	X	X	X	X	X	X
RNTN	0.287	0.264	X	X	X	X	X
CAR	0.0167	0.0355	0.0214	X	X	X	X
VIX	-0.0478	-0.0439	-0.037	-0.0519	X	X	X
Length	-0.2533	0.1998	0.2034	0.0514	-0.0092	X	X
ROE	0.185	0.1483	0.0339	0.028	-0.027	-0.0825	X
B/M	-0.044	-0.0533	0.0089	-0.0052	0.0106	0.0598	0.1633

higher. Looking at the graphs you can see that during the highlighted period which are stock market crashes the sentiment for all classifiers goes down, to increase again afterwards. All models see a dip in sentiment for the dot.com crash, financial crisis and the corona crisis, however the Finbert has a stronger difference between the boom and bust times.

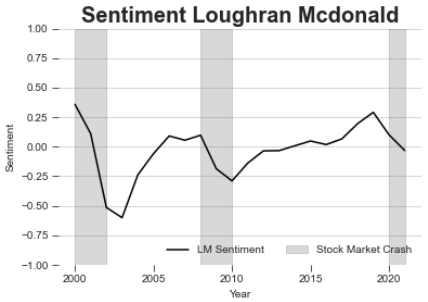


Figure 2: Sentiment Loughran McDonald period 1999-2021

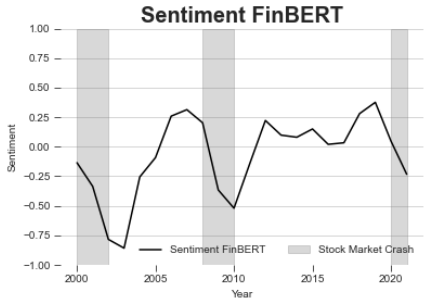


Figure 3: Sentiment FinBERT period 1999-2021

4 RESULTS

4.1 *Financial Phrasesbank and FIQA*

The results of the first sentiment classification experiment are presented in table 3 and visualized in figure 4 and 5. The results provide the answer to the first research question; To what extent can the level of syntactic information the models are capable of representing improve the accuracy of the sentiment classification of labeled financial sentences? The dictionary-based methods use no level of syntactic information and therefore you would expect that they have the most difficulty with correctly classifying the sentiment. The Harvard IV-4 dictionary achieves an accuracy of 0.376

and F1-score of 0.365 for the Financial Phrasebank including all sentences with at least an agreement rate of the 16 experts of at least 50%. The Loughran and McDonald the other dictionary-based method achieves a significantly higher accuracy of 0.575 and F1-score of 0.570. The Loughran and McDonald dictionary outperforms the Harvard IV-4 for all the different levels of agreement. Li (2008) and Loughran and McDonald (2011) criticized the use of non-financial dictionaries to capture the sentiment for financial and accounting language. The result I find supports this critique that using non-financial based dictionaries lead to worse results due to the specific nature and different meaning of words in a financial or accounting context. The third model FinBERT achieves the highest score of the models, an accuracy for the 50% agreement dataset of 0.773 and a F1-score of 0.761. The accuracy is 0.198 higher than the second-best model and the F1-score is 0.191 higher than the second-best model. The FinBERT model incorporates word order as syntactic information and is pre-trained and fine-tuned on financial literature. The RNTN considers compositionality as syntactic information and has an accuracy of 0.492 and F1-score of 0.493. The RNTN model outperforms the non-financial dictionary but fails to outperform the Loughran and McDonald dictionary. When Socher et al. (2013) published his work on the RNTN the model outperformed the state-of the art models of his time on a variety of tasks. Therefore, I expected a higher performance than the dictionary-based methods; this is not the case for the Financial Phrasebank. The RNTN model from the Stanford CoreNLP uses the Stanford Treebank which consists of movie reviews. Which might be the reason the model does not outperform a model specially created for financial literature such as the Loughran and McDonald dictionary.

Table 3: Model performance on Financial Phrase Bank, FIQA and MDA dataset.

Models	Harvard IV-4		LM		FinBERT		RNTN	
Dataset			Accuracy		F_1-score			
Phrasebank								
100% Agree	0.367	0.365	0.602	0.594	0.89	0.885	0.516	0.516
75% Agree	0.366	0.360	0.60	0.592	0.852	0.846	0.514	0.514
66% Agree	0.373	0.359	0.583	0.577	0.808	0.799	0.498	0.499
50% Agree	0.376	0.360	0.575	0.570	0.773	0.761	0.492	0.493
FIQA								
Headlines	0.397	0.449	0.635	0.664	0.642	0.676	0.667	0.680
MicroBlogs	0.394	0.427	0.533	0.545	0.575	0.587	0.496	0.517
10-Ks								
MDA Section	0.227	0.225	0.488	0.519	0.783	0.780	0.428	0.475

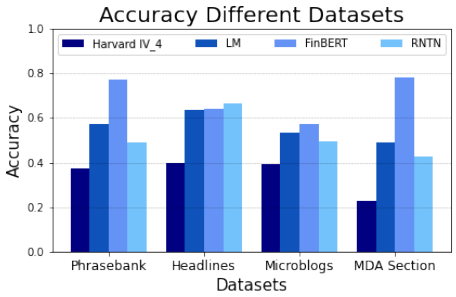


Figure 4: Accuracy of the four models on the datasets

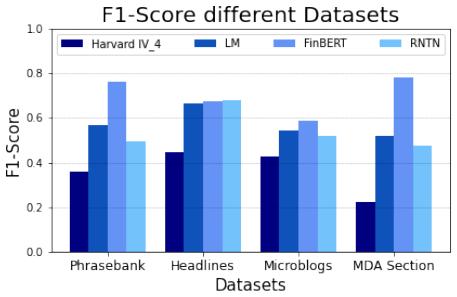


Figure 5: F1-score of the four models on the datasets

The second dataset in table 3 is the FIQA dataset which consists of financial headlines and microblogs: posts on stock twits, reddit and twitter. The overall results are similar, where the Harvard IV-4 performs the worst on both dataset with an accuracy of 0.397 and F1-score of 0.449 for the headlines and an accuracy of 0.394 and 0.427 for the microblogs which are comparable to results on the Financial Phrasebank. The Loughran and McDonald performance is also similar to the performance for Financial Phrasebank with an accuracy of 0.635 and F1-score of 0.664. The FinBERT model however performance is significantly lower on both datasets, and for the headlines no longer the best performing model. The RNTN model's performance is highest for the headlines. The RNTN model sees compared to the LM Dictionary and FinBERT model no drop in the performance on the tweets compared to the performance on the Financial Phrasebank. The reason for the smaller difference in comparison with the FinBERT and the LM dictionary on the microblogs might be due to the similarity between movie reviews and microblogs

The final dataset MDA section in table 3 is annotated by myself due to the lack of the availability labeled data and the need to measure the performance on different forms of financial texts. The performance of Harvard IV-4 is 0.227 and 0.225 for the accuracy and the F1-score. The model performance significantly worse on the dataset consisting of sentences from corporate disclosures then on the other two datasets. The Loughran and McDonald dictionary has an accuracy of 0.488 and 0.519 which is lower than the other two datasets but is significantly higher than the non-financial dictionary. The FinBERT model has an accuracy of 0.783 and F1-score of 0.78. The model performs similarly to the other two datasets and significantly outperforms all other methods with a difference to the second best model on the MDA section dataset of 0.295 for the accuracy and 0.261 for the F1-score. The RNTN model has an accuracy of 0.428 and

F1-score of 0.475, which is consistent with the performance on the other two datasets and to the Loughran and McDonald dictionary.

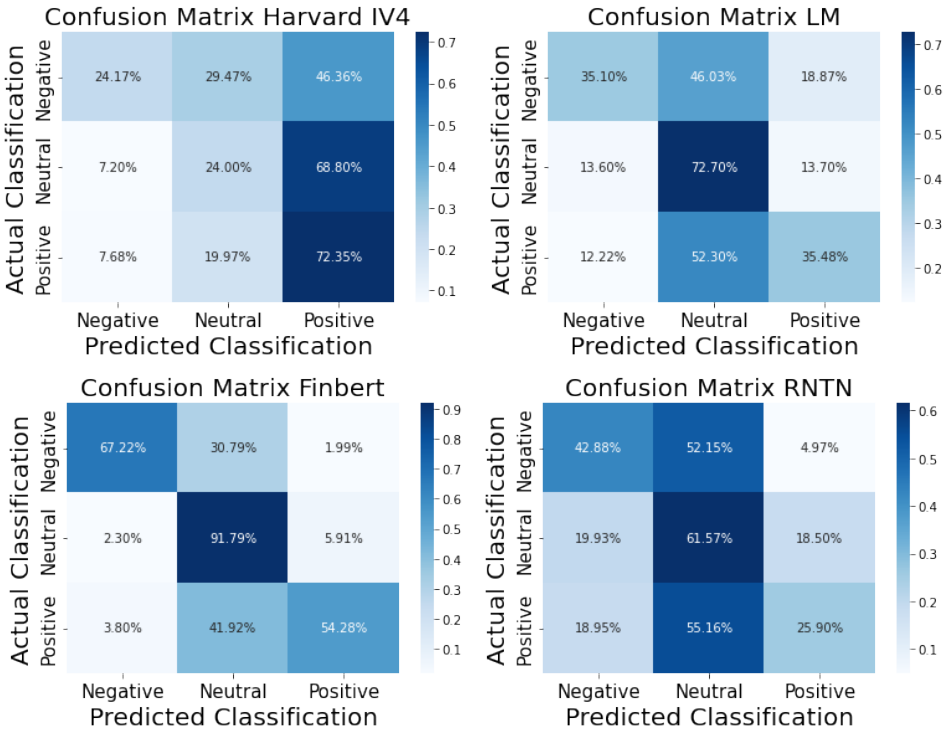


Figure 6: Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from Financial Phrasebank 50% agreement.

In figure 6 you see the confusion matrix of all the four models for the Financial Phrasebank 50% agreement dataset. The superior performance of the FinBERT stands out especially when correctly predicting neutral, and negative sentiments. Surprisingly the Harvard IV-4 model achieves the highest score on correctly predicting positive sentiments. However, the model achieves a very low score on the correctly predicting of the negative and neutral sentiments indicating that it predicts a positive sentiment way too often. The LM dictionary and the RNTN model are both fairly good at correctly predicting the negative sentiment but seem to have a problem with predicting the positive sentiment correctly. The confusion matrices of the other levels of agreement can be found in the appendix.

In figure 7 you see the confusion matrix for all the four models for the FIQA headline dataset. The FinBERT model is also the best performing model in this confusion matrix however the difference is much smaller than for the Financial Phrasebank. The FinBERT has the highest performance in it correctly classifying negative sentiments and the positive sentiment. The RNTN has the highest score for the neutral segment. Furthermore,

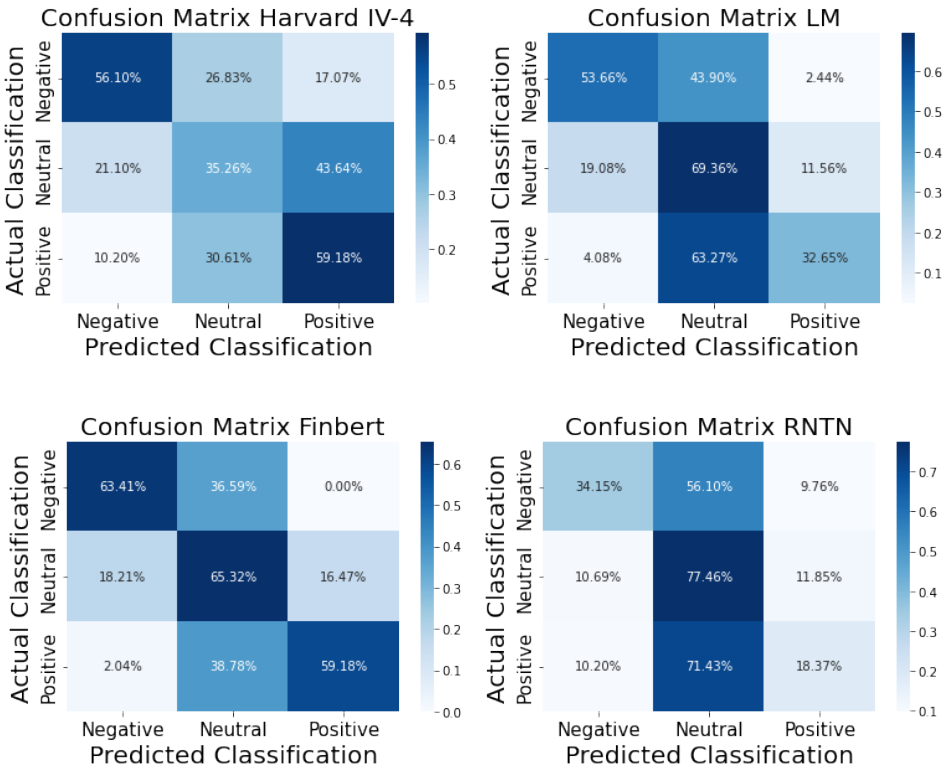


Figure 7: Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the FIQA headline dataset.

the performance of the RNTN model on the positive sentiment stands out where the model only has an accuracy of 18.37%. The Harvard IV-4 is also in this dataset relative good at correctly classifying the positive segment where it has the highest score of 59.18% together with the FinBERT model

Overall, the FinBERT is the best performing model for sentiment classifiers and the LM dictionary is the second-best performing model. The difference with the RNTN is not very large but especially for the Financial Phrasebank significantly. In the section below I will present several errors made by the models and highlight some frequently occurring mistakes.

Error 1: *astrazeneca chases acerta to secure next cancer drug winner.*
***LM**: Neutral **Harvard**: Neutral **FinBERT**: Positive **RNTN**: Positive*

For error 1 the positive sentiment classification is the correct one; the Harvard IV-4 and the LM dictionary both fail to correctly classify the sentence as where the FinBERT and RNTN model correctly classify the sentiment. This error is an example of the disadvantage of using dictionary based-methods highlighted in section 2.1. Specific words belonging to a particular industry in this case the pharmaceutical industry can have a

different meaning when it is applied in a 10-k report. The Harvard IV-4 considers the word cancer as negative and the word winner as positive leading to a classification of neutral. While securing a new cancer drug winner is very positive for AstraZeneca. Both the FinBERT and the RNTN are able to capture the correct sentiment, where this for the FinBERT which is trained on financial literature could be expected. However the RNTN is based on movie reviews, but still is able to capture the correct sentiment. The LM dictionary classifies the sentence wrong but it is not because it considers the word cancer negative, but because none of the words in the sentence are considered to be positive or negative. The advantage of the LM dictionary is that the dictionary is tailored to financial language, however this can also lead to errors due to the strict selection of words which are included.

Error 2: *slump in weir leads ftse down from record high.*
***LM**: Positive **Harvard**: Neutral **FinBERT**: Negative **RNTN**: Negative*

For error 2 the negative sentiment is the correct classification. The LM method classified the sentence as positive and the Harvard IV-4 classified the sentence as neutral. The dictionary methods do not take into account the position of words, while the other methods do. Therefore the dictionary models do not capture that the index goes down coming from a record high, while both the FinBERT and RNTN are able to capture the correct sentiment.

Error 3: *amzn new hod with conviction keeping on watch for resistance.*
***LM**: Negative **Harvard**: Neutral **FinBERT**: Positive **RNTN**: Neutral*

For error 3 the correct sentiment is positive, only the FinBERT model was able to correctly predict the sentiment. The last error is a sentence from social media and uses complex language. The abbreviation 'hod' which stands for high of the day is a form of informal language. Dictionary models have difficulty with sentences that include informal language because they do not include abbreviations. Which is understandable when classifying sentiment from corporate filings due to the infrequent use of abbreviations, however when classifying social media posts this leads to problems due to the frequent use of abbreviations. The RNTN model classifies the sentence as neutral, although it is not possible to backtrack why exactly, potential reason could be the lack of understanding of financial language such as 'hod'. The FinBERT model clearly demonstrates that it is able to understand complex and informal language from social media as well the formal language from corporate filings.

4.2 *Event Study: Management Discussion and Analysis*

The second research question intends to answer if there is a relationship between the CARS and the sentiment extracted from the MDA section of a 10-k report. In table 4 you can find the OLS regression. In the first regression model only the dependent variable CAR[-1,1] and the independent variable sentiment LM is included. The estimate of the sentiment LM is positive but only significant on the lowest significance level. The coefficient is 0.008 indicating that an increase in the sentiment of one leads to an increase in the CAR of 0.8%. The standard deviation is 0.0002. When management is positive the stock returns are higher than when management is pessimistic. However when adding the control variables in model 4. The LM sentiment is no longer significant; this indicates that the LM sentiment classifier does not have a relationship with the CAR.

Table 4: Estimation of the regression equation for the sentiments; depedent variable CAR[-1,1]

Models	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0	0.0003	0.001	0.006	0.0023	0.0028
	(0)	(0)	(0)	(0.001)	(0.001)	(0.006)
LM	0.008*			0.002		
	(0.0002)			(0.001)		
FinBERT		0.001**			0.001 *	
		(0.0003)			(0)	
RNTN			0.005**			0.006**
			(0)			(0)
b/m				0	0.001	0
				(0)	(0)	(0)
ROE				0.0018***	0.0021**	0.0043***
				(0.001)	(0.001)	(0.001)
VIX				-0.0002	0.0009*	0
				(0)	(0.0001)	(0)
Length				0***	0	0*
				(0)	(0)	(0)
Obs	4988	4988	4988	4416	4416	4416
R2	0.001	0.001	0.001	0.006	0.003	0.007
Adj. R2	0	0.001	0.001	0.005	0.002	0.008
F-statistic	3.293	6.33	3.945	5.294	3.209	6.179

Note: The table includes the regression coefficient and the T-statistic; *p<0.1; ** p<0.05; ***p<0.01

The second model consists of only the dependent variable and the FinBERT sentiment as the independent variable. The FinBERT coefficient

is positive indicating that companies with a more positive sentiment in the MDA section have higher CARs. The coefficient is 0.001 indicating that an increase in the sentiment of 1 leads to a CAR of 0.1%. When adding the control variables in model five the FinBERT sentiment is still significant however on a lower significance. The FinBERT sentiment is positive indicating that a higher sentiment leads to larger CARs. This is in line with my expectations because due to the position of the management they have an information advantage so when management is more positive this is interpreted by the market as a positive signal. Furthermore, is the ROE significant and the VIX, companies with a higher ROE have higher CARs, similar for the VIX a larger VIX has a positive effect on the CAR which is also in line with the expectation that when the volatility is higher the market response to information is more dispersed. The overall model has 4416 observations and an r-squared of only 0.003 indicating that there are other variables who have an effect on the CARs surrounding the release of the 10-k.

The last sentiment classifier RNTN is significant in model 3 when you do not include control variables. The RNTN is positive indicating that a company with a more positive sentiment score will have higher CARs. The coefficient is 0.005 indicating that an increase of the sentiments of 1 will lead to an increase of 0.5&. When adding the control variables in model 6 in table 4 the RNTN classifier is still significant and has a coefficient of 0.006. The RNTN model has a higher significance level than the FinBERT classifier. Furthermore for the control variables especially the ROE has a significant positive relationship with the CARs and the text length is also positive and significant buth the effect is very small. The model has an adjusted r-squared of 0.007 which is larger then for the model 5 of FinBERT but is still very low.

5 DISCUSSION

The aim of the paper is to compare the methods traditionally used to classify sentiment in finance and accounting research with more advanced NLP techniques which incorporate syntactic information. The best performing models on the labeled data are then used to understand if you can use the sentiments extracted from the MDA section of a 10-k report to forecast stock prices. Below you find the answer to the first research question:

RQ1 *To what extent can the level of syntactic information the models are capable of representing improve the accuracy of the sentiment classification of labeled financial sentences?*

The results indicate that advanced NLP techniques which incorporate some form of syntactic information can increase the performance compared to non-financial dictionaries. For all the labeled datasets the advanced NLP techniques outperform the Harvard IV-4 dictionary on accuracy and F1-score. However, the use of syntactic information is no silver bullet. When comparing the advanced NLP techniques to a dictionary designed especially for financial language the advanced models do no longer outperform on all labeled datasets. The RNTN model which is not designed for financial language fails to outperform the LM dictionary on all but one section of the labeled data. The FinBERT model which is pre-trained and fine-tuned on financial language is the only advanced model which outperforms both dictionaries non-financial and financial. The study concludes that the use of some form of syntactic information leads to better performance but can not make up for the specific nature of financial language. Ideally incorporating some form of syntactic information is combined by designing the model using financial texts such as corporate filings.

The second research question is as follows:

RQ2 *: Does the sentiment extracted from the MDA section of 10-k reports have a significant positive relationship with the abnormal returns surrounding the release of a 10-k?*

The study finds that the sentiment classifiers of the advanced NLP techniques have a significant relationship with the CARs. This indicates that there is a relationship between the sentiment in the MDA section of a 10-k report. The event study was done using the three best performing models from the previous research question. The LM dictionary, RNTN and FinBERT model were used to classify the sentiment. The LM dictionary was the only model which was not significant when adding the control variables. THe RNTN and FinBERT were significant; the results indicated a positive relationship. So the sentiment extracted from the MDA section of a 10-k report has a significant relationship with the abnormal returns surrounding the release of a 10-k. However the results are small with a positive effect of increasing the sentiment with one of 0.5% for the RNTN and 0.1% of the FinBERT. If you would want to use the sentiment classifiers to make trading decisions on the stock market you would not be able to translate the relationship into positive returns due to factors such as transaction fees and other market frictions. It therefore might be an opportunity to search for an event of release of information which has a larger effect on the stock prices so that the effects will be larger enough to overcome the constraints of transaction fees. Potentially interesting events could be quarterly earnings or meetings of central banks which do have a large effect on the financial markets.

5.1 *Contributions, limitations and suggestions for future research*

In this thesis several approaches to classifying the sentiment of financial language are modeled and compared. The models included traditional used methods such as the dictionary approach and advanced NLP techniques such as a BERT and an RNTN model. The use of advanced NLP techniques can lead to better performance on the accuracy and F1-score of the models. The study shows that despite the specialized knowledge necessary for using the advanced NLP techniques they have advantages and can lead to significant improvement compared to the traditional methods. The study encourages the use of the advanced NLP compared to the traditional methods. Furthermore, the study shows that when applying advanced NLP techniques the best results are obtained by designing them using financial language.The study also shows that sentiment classifiers of the MDA section of a 10-k have a significant positive relationship with the abnormal returns surrounding the release of the information. This would indicate that you can use the sentiment classifiers to forecast stock prices. However due to factors such as transaction fees this will not result in a positive return compared to investing in the index.

The most serious limitation of this study is the lack of available annotated data on corporate filings. The labeled datasets available consist of sentences from news articles and social media posts from stock forums. This creates questions about the generalization of the four models on corporate filings. This study tries to overcome this problem by annotating 600 sentences from the MDA section of 10-ks. The classification of financial literature is however complicated and requires specialized knowledge. Despite having a background in finance myself, the disagreement between the experts who annotated the Financial Phrase bank shows the need for a group of experts instead of a single person to establish ground truths.

Another limitation is that not all models which are compared are built using financial language. The Harvard IV-4 dictionary is built to be a general purpose dictionary and the RNTN model is built using movie reviews. Whereas the LM dictionary and the FinBERT model are both especially designed to be used for financial language. To better understand to what effect the improvement in the performance is due to the use of syntactic information and what is due the special nature of financial language, future research could build a treebank for the RNTN model based on financial literature and especially corporate filings.

Furthermore, due to lack of available time and constraint on the computer power only the MDA section of the 10-k was used to determine a sentiment score. The use of a small subset of the 10-k might not be a good proxy for the sentiment of the company/management of the entire

document. The ability to include a large part of the 10-k might lead to a better representation of the sentiment from the company. It could also give insight into the level of information in each individual section.

In addition to the potential solution offered to the limitations mentioned above, another suggestion for future research is to use the advanced NLP techniques mentioned above, especially FinBERT for fraud detection. The FinBERT model outperforms the other models on the labeled dataset and is the only model which is able to extract a sentiment score which is significant with the CARs. Although the results are too small to use in the real world this might differ for fraud detection as fraud in corporations tends to have severe negative consequences.

6 CONCLUSION

In this thesis several approaches to classifying the sentiment of financial language are modeled and compared. The models included traditional used methods such as the dictionary approach and advanced NLP techniques such as a BERT and an RNTN model. The use of advanced NLP techniques can lead to better performance on the accuracy and F1-score of the models. The study shows that despite the specialized knowledge necessary for using the advanced NLP techniques they have advantages and can lead to significant improvement compared to the traditional methods. The study encourages the use of the advanced NLP compared to the traditional methods. Furthermore, the study shows that when applying advanced NLP techniques the best results are obtained by designing them using financial language.The study also shows that using sentiment classifiers to forecast stock prices is possible, but that the potential results are too small to create a trading strategy due to constraints such as transaction fees.

REFERENCES

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, *abs/1908.10063*. Retrieved from http://arxiv.org/abs/1908.10063

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Cable, J., & Holland, K. (1999). Modelling normal returns in event studies: a model-selection approach and pilot study. *The European Journal of Finance*, *5*(4), 331–341.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2020, October). *How to talk when a machine is listening?: Corporate disclosure in the age of ai* (Working Paper No. 27950). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w27950 doi: 10.3386/w27950

Das, S. R., & Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management science*, *53*(9), 1375–1388.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383–417.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915–953.

Field, L., Lowry, M., & Shu, S. (2005). Does disclosure deter or trigger litigation? *Journal of Accounting and Economics*, *39*(3), 487–507.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, *45*(4), 363–407.

Henry, E., & Leone, A. J. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review*, *91*(1), 153–178.

Hoberg, G., & Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, *43*, 58–85.

Huang, A., Wang, H., & Yang, Y. (2020). Finbert—a deep learning approach to extracting textual information. *Available at SSRN 3910214*.

Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185.

Kothari, S. P., Shu, S., & Wysocki, P. D. (2009). Do managers withhold bad news? *Journal of Accounting research*, *47*(1), 241–276.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *CoRR*, *abs/1710.03954*. Retrieved from http://arxiv.org/abs/1710.03954

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, *45*(2-3), 221–247.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance*, *16*(1), 1–11.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance:: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, *12*, 357–375.

Loukas, L., Fergadiotis, M., Androutsopoulos, I., & Malakasiotis, P. (2021, November). EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the third workshop on economics and natural language processing* (pp. 13–18). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.econlp-1.2

Lutz, B., Pröllochs, N., & Neumann, D. (2019). *Sentence-level sentiment analysis of financial news using distributed text representations and multi-instance learning.* arXiv. Retrieved from https://arxiv.org/abs/1901.00400 doi: 10.48550/ARXIV.1901.00400

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, *35*(1), 13–39. Retrieved 2022-06-16, from http://www.jstor.org/stable/2729691

Maia, M., Freitas, A., & Handschuh, S. (2018). Finsslx: A sentiment analysis model for the financial domain using text simplification. In *2018 ieee 12th international conference on semantic computing (icsc)* (p. 318-319). doi: 10.1109/ICSC.2018.00065

Malo, P., Sinha, A., Takala, P., Korhonen, P., & Wallenius, J. (2013). *Good debt or bad debt: Detecting semantic orientations in economic texts.* arXiv. Retrieved from https://arxiv.org/abs/1307.5336 doi: 10.48550/ARXIV.1307.5336

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012a). Earnings conference calls and stock returns: The incremental informativeness

of textual tone. *Journal of Banking & Finance*, *36*(4), 992–1011.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012b). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, *36*(4), 992-1011. Retrieved from https://ideas.repec.org/a/eee/jbfina/v36y2012i4p992-1011.html doi: 10.1016/j.jbankfin.2011.1

Purda, L., & Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, *32*(3), 1193–1223.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training (2018).*

ROLL, R. (1988). R2. *The Journal of Finance*, *43*(3), 541-566. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1988.tb04591.x doi: https://doi.org/10.1111/j.1540-6261.1988.tb04591.x

Shiller, R. J. (1980). *Do stock prices move too much to be justified by subsequent changes in dividends?* National Bureau of Economic Research Cambridge, Mass., USA.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642).

Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2017). Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, *5*, 1-25.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

7 APPENDIX A

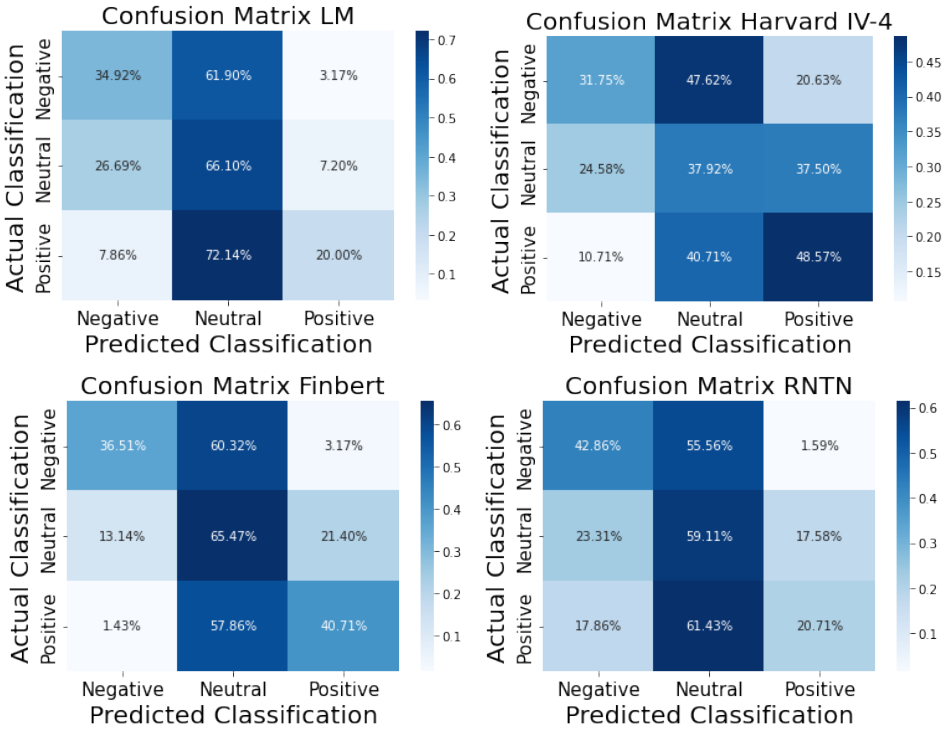


Figure 8: Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the FIQA microblogs dataset.

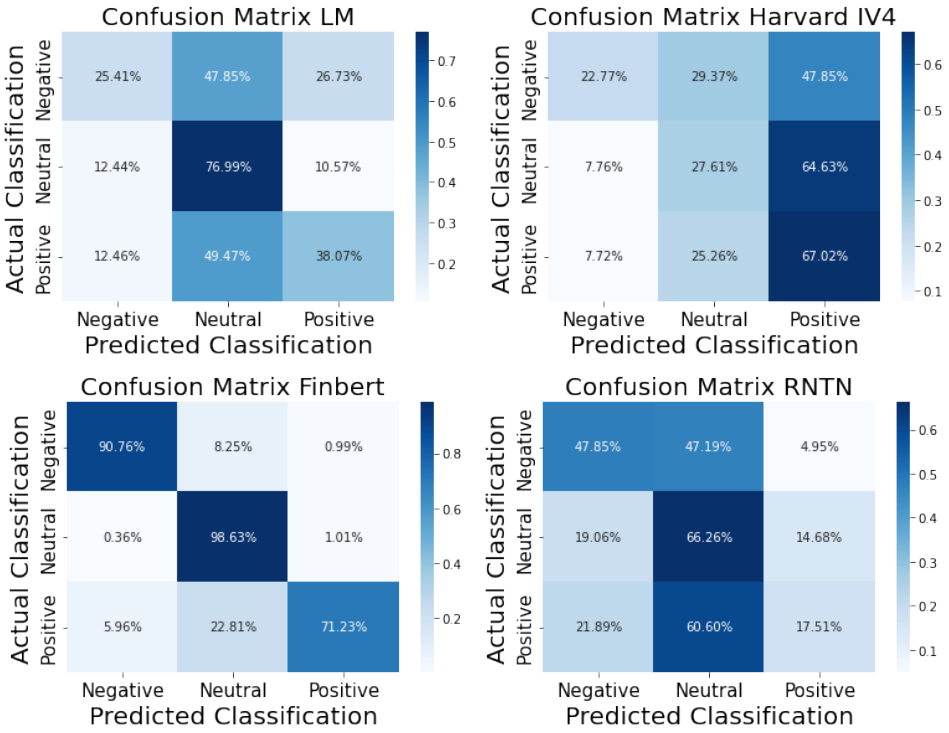


Figure 9: Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the Financial Phrasebank 100% agreement.

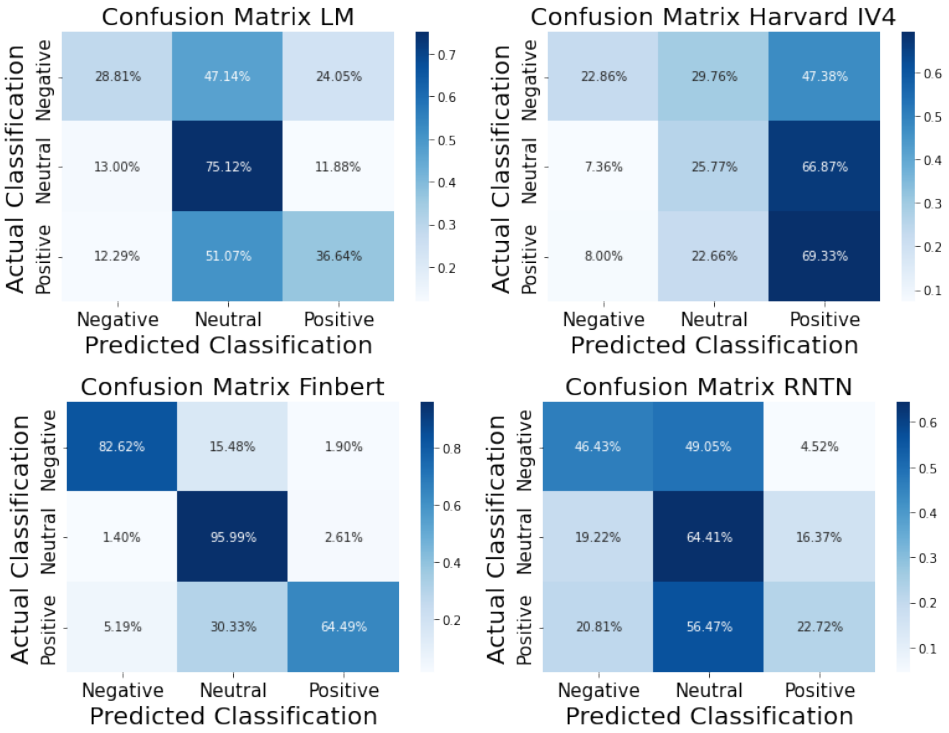


Figure 10: Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the Financial Phrasebank 75% agreement.

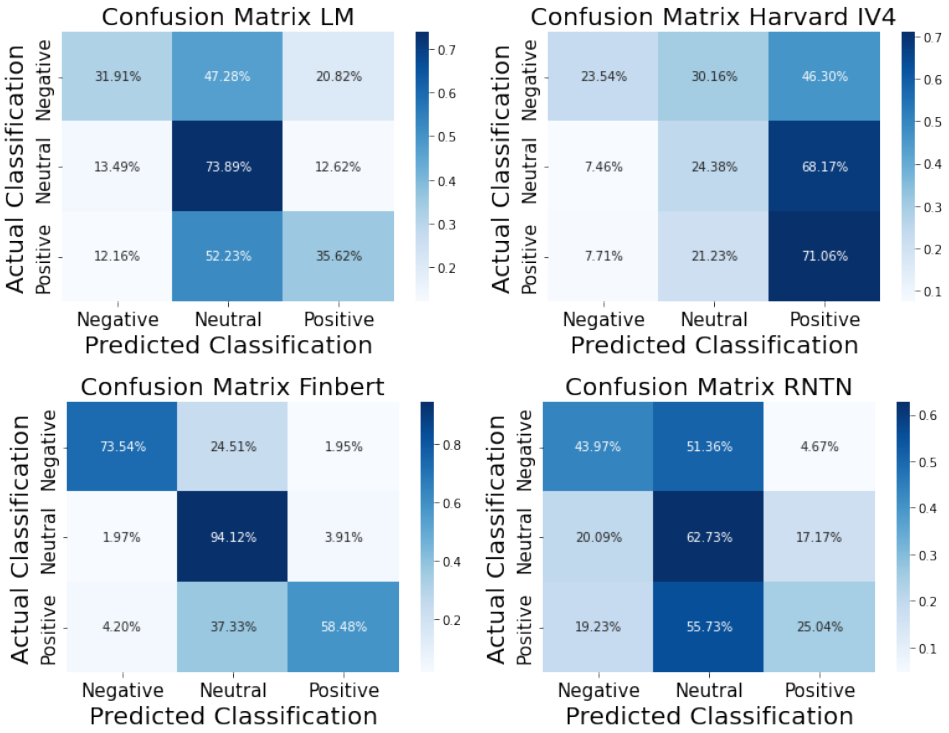


Figure 11: Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the Financial Phrasebank 66% agreement.

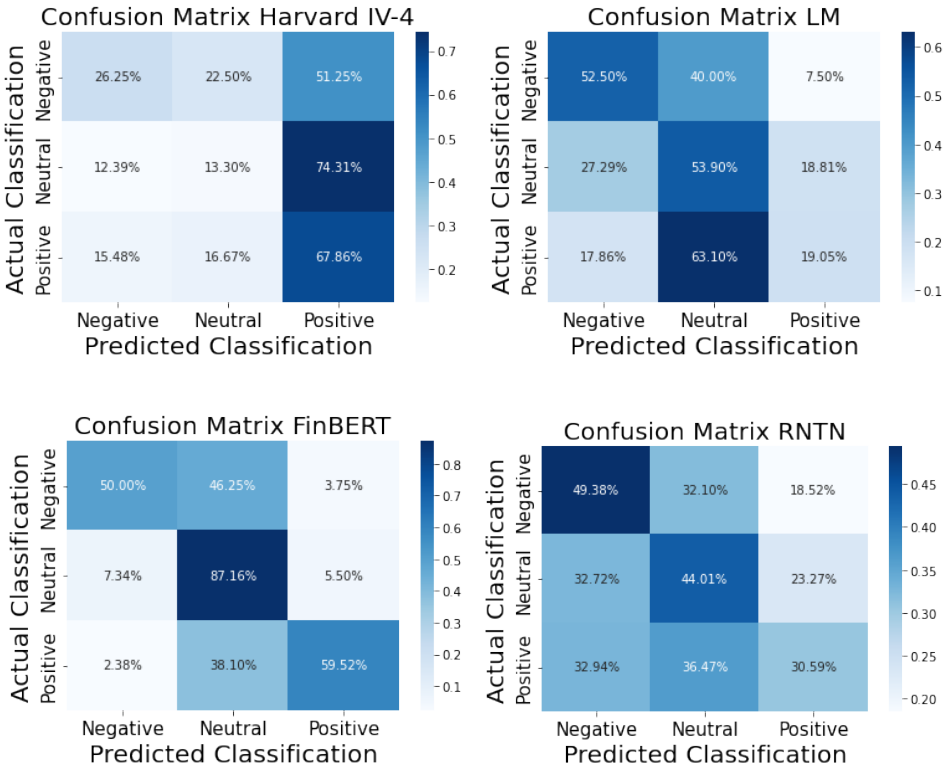


Figure 12: Confusion Matrix of the Harvard IV-4, LM, FinBERT and RNTN from the MDA section dataset

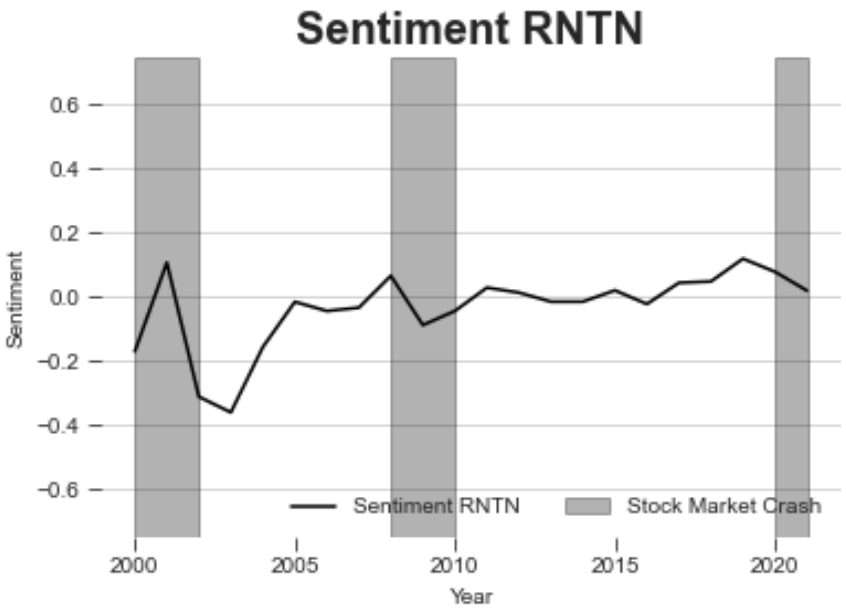


Figure 13: Sentiment RNTN period 1999-2021